DOCUMENT RESUME

ED 482 099 IR 058 787

AUTHOR	Shabajee, Paul; Miller, Libby; Dingley, Andy
TITLE	Adding Value to Large Multimedia Collections through Annotation Technologies and Tools: Serving Communities of Interest.
PUB DATE	2002-04-00
NOTE	15p.; In: Museums and the Web 2002: Selected Papers from an International Conference (6th, Boston, MA, April 17-20, 2002); see IR 058 778.
AVAILABLE FROM	For full text: http://www.archimuse.com/mw2002/papers/shabajee/shabajee.html/.
PUB TYPE	Reports - Descriptive (141) -- Speeches/Meeting Papers (150)
EDRS PRICE	EDRS Price MF01/PC01 Plus Postage.
DESCRIPTORS	*Access to Information; Biodiversity; Databases; Foreign Countries; *Indexing; Information Retrieval; Information Technology; *Metadata; Models; *Multimedia Materials; World Wide Web
IDENTIFIERS	Interoperability; Ontology; Semantic Webbing; United Kingdom

ABSTRACT

A group of research projects based at HP-Labs Bristol, the University of Bristol (England) and ARKive (a new large multimedia database project focused on the world's biodiversity based in the United Kingdom) are working to develop a flexible model for the indexing of multimedia collections that allows users to annotate content utilizing extensible controlled vocabularies. As part of the educationally focused ARKive-ERA project, a series of models for user annotation have been developed. One example is that of university lecturers and researchers studying a particular type of animal behavior. They may wish to identify all relevant images or video of that particular behavior and annotate them as good illustrations of aspects of that behavior. However, significant issues arise over, for example, the validation of information, access control and the use of such annotations by the resource discovery tools. The paper explores these and other issues and problems involved, and explains how the various models can help provide solutions to key problems and thus meet the needs of a diverse range of communities of interest, thereby adding significant value to online multimedia collections. (Contains 22 references.) (Author/MES)

Reproductions supplied by EDRS are the best that can be made
from the original document.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

D. Bearman

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

PAPERS
Museums and the Web 2002

Adding Value To Large Multimedia Collections Through Annotation Technologies And Tools: Serving Communities Of Interest

Paul Shabajee, Libby Miller, Institute for Learning and Research Technology (ILRT), University of Bristol, UK, Andy Dingley, Codesmiths, UK

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

ED 482 099

IR058787

Register
Workshops
Sessions
Speakers
Interactions
Demonstrations
Exhibits
Events
Best of the Web
Key Dates
Boston
Sponsors

A&MI

Archives & Museum Informatics
2008 Murray Ave.
Suite D
Pittsburgh, PA
15217 USA
info@archimuse.com
www.archimuse.com

Search A&MI

Join our Mailing List.
Privacy.

Abstract

A group of research projects based at HP-Labs Bristol, the University of Bristol and ARKive (a new large multimedia database project focused on the world's biodiversity based in the UK) are working to develop a flexible model for the indexing of multimedia collections that allows users to 'annotate' content utilizing extensible controlled vocabularies. As part of the educationally focused ARKive-ERA project, a series of models for user 'annotation' have been developed.

The need for these types of user support and tools was identified while conducting pre-design user studies with specialist user groups. The needs center around the limitations of current on-line museum and library systems that do not provide support for users to annotate or 'tag' multimedia objects of relevance to their particular 'community of interest' or with specialized indexing terms. Tagging would enable specialized resource discovery and knowledge sharing with other members of their communities.

One example is that of University Lecturers and Researchers studying a particular type of animal behavior. They may wish to identify all relevant images or video of that particular behavior and annotate them as good illustrations of aspects of that behavior. However, significant issues arise over, for example, the validation of information, access control and the use of such annotations by the resource discovery tools. The paper explores these and other issues and problems involved, and explains how the various models can help provide solutions to key problems and thus meet the needs of a diverse range of 'communities of interest', thereby adding significant value to on-line multimedia collections.

Key Words: community annotation, flexible publishing, semantic web, ontologies, collaboration

Introduction

The ARKive-ERA project is focused on investigating how best to design the underlying technological infrastructures to enable large multimedia database systems to maximize the educational potential of their multimedia assets, for users from very diverse range of backgrounds and in a wide variety of contexts. The focus for the research has been the ARKive project (http://www.wildscreen.org.uk/arkive/), a large multimedia Web-based database system under development, containing diverse data related to endangered animal, plant and fungi species and their habitats as well as more common UK species.

BEST COPY AVAILABLE

ARKive is characteristic of many large digitization projects; during its *initial phase* of development it will contain data profiling some 2000 species and their habitats. This will take the form of approximately 9,000 minutes of digitized video and 30,000 still images along with hours of audio, maps, textual information and other supporting media and educational materials. These assets are donated by aa diverse range of commercial and non-profit organizations as well as by individuals.

Essentially ARKive is a 'community project' insofar as it is part of and relies on a community of organizations and individuals who have an interest in sharing access to rich multimedia resources focused on biodiversity.

ARKive type projects are designed to serve the needs of their diverse potential users by providing tools for individuals and communities of users to 'annotate' the content of the database so as to make the content more valuable to others with similar interests.

It is important to note that we define *annotation* as metadata (see below) created after the creation of the content. It is this post hoc nature ("a note added to anything written", Oxford English Dictionary, 1998) that represents a considerable expansion of its usefulness, as a means of adding value to content, because it now allows people other than the original content author to add metadata descriptions.

The Challenge

As part of early work to identify the key requirements of ARKive and similar projects, it became clear that the diverse range of potential users includes school children and their teachers, media researchers, conservation scientists, customs officers, university lecturers and students and the very many people with personal rather than professional or educational interests in wildlife, to name but a few.

Each of the groups listed above is itself diverse with respect to the particular needs or desires of ARKive or similar systems.

We conducted a small-scale interview survey with University Lecturers about their likely uses and needs of ARKive with respect to supporting their teaching activities. As a result we identified that a key need of this 'group of users' was to be able to search for multimedia resources to 'illustrate' concepts when presenting to and supporting their students. Specific examples included infanticide, drug induced behavior, life strategies of plants, inter-specific competition, identifying and classifying organisms, tropic levels, echolocation, binocular vision, and harvesting theory. Lecturers from different sub-domains (e.g. behavioral biology and ecology) suggested different terms.

This small example shows that even within sub-groups of one relatively well-defined group of users the 'resource discovery needs' alone were complex. Indeed it quickly became clear that all of these sub-groups or 'communities of interest' have their own specialized vocabularies and concepts that they would like the resources indexed under so as to support their resource discovery needs.

Not only did they want to be able to search for assets using specialist vocabularies, but they also wanted ideally be able to find 'good' examples of assets that illustrated a particular aspect of a particular concept.

These are not unreasonable requests as clearly it is not practical to browse 9,000 minutes of video, or even a small sub-set, in the hope of finding a good example to illustrate aspects of, for example, 'harvesting theory'. However for ARKive it is simply not feasible to index every asset with the terms relevant to all possible communities of interest.

The problems can be expressed more clearly by using basic concepts from Information Retrieval (IR) literature (e.g. Chowdhury 1999).

- **Precision** (number of relevant documents in results divided by the number of retrieved objects)
- **Recall** (number of retrieved documents divided by the number of relevant objects in the collection)
- **Specificity** (level of detail of indexing of an object i.e. how fine grained is the indexing)
- **Exhaustivity** (completeness of indexing of object using available vocabulary)

If we use the example above, the University lectures (not unreasonably) want to use their specialist vocabularies to search ARKive's multimedia archive and have high precision and recall from the system based on those terms and queries constructed from them.

When actually doing the indexing, ARKive has to balance the levels of specificity and exhaustiveness of their indexing to make the task tractable within the limits of available resources and time.

It is useful here to refer to some well-defined 'specialist vocabularies' to get some insight into the scale of the challenge.

- The Biosys, Zoological Record (http://www.biosis.org.uk/products_services/zoorecord.html) is indexed using an extensive thesaurus of around 10,000 terms developed over more than 20 years.
- The General Multilingual Environmental Thesaurus (GEMET) was developed by the European Environment Agency (EEA) together with a co-operation of international experts to serve the needs of environmental information systems. Analysis and evaluation work produced a core terminology of 5,400 generalized environmental terms and their definitions. (European Environment Agency, 2002)
- Medical Subject Headings (MeSH) (http://www.nlm.nih.gov/mesh/meshhome.html)contain more than 19,000 main headings. There are also 103,500 headings called Supplementary Concept Records within a separate chemical thesaurus.
- The Art & Architecture Thesaurus (AAT - http://www.getty.edu/research/tools/vocabulary/aat/) is maintained by the Getty Research Institute (see http://www.getty.edu/gri/) and the thesaurus keeps growing. The AAT contains about 120,000 terms covering objects, textual materials, images, architecture and material culture from antiquity to the present. (J. Paul Getty Trust, 2002)
- UK National Curriculum (for schools) Metadata Schema contains some 2000 'subject keywords' (Qualifications and Curriculum Authority, 2002)

While any particular collection of multimedia will not necessarily contain objects which are appropriately indexed under many of the terms from each and every available specialist vocabulary, it is none-the-less possible that

some terms from all of these will be applicable to some objects.

These issues are relevant to many other types of projects, not least those involved in the development of cross-searching of multiple databases which have been indexed using different indexing schemas (e.g. Clark, 2001). Much work is going on with respect to providing ways of robustly mapping between different schemas. These issues are discussed again below.

Indexing, Metadata, Interoperability and Ontologies

It is beyond the scope of this paper to review the extensive literature on the indexing and applications of 'metadata' (see below) to multimedia objects, the issues of interoperability and related technologies. See Gill and Miller (2002) for an overview of the key issues with regard to 'digital cultural content', and below for some examples of relevant projects. However a brief overview is necessary and useful in providing additional background to the remainder of the paper.

Metadata is broadly defined as 'data about data ' (Gilliland-Swetland, 1998). The traditional library catalogue index card is a classic example of metadata. The publication date, author, title, publisher, dewey decimal code... are *'metadata elements'* within a clearly defined *metadata schema* and *scheme* (list of metadata elements, allowed states of those and relationships between them).

As can be seen from this example, there are different types of metadata. Gilliland-Swetland (1998) distinguishes between 5 types:

- **Administrative:** Metadata used in managing and administering information resources
- **Descriptive:** Metadata used to describe or identify information resources
- **Preservation:** Metadata related to the preservation management of information resources
- **Technical:** Metadata related to how a system functions or metadata behave
- **Use:** Metadata related to the level and type of use of information resources

Descriptive metadata is of most relevance to the challenges outlined above, but in principle the issues apply to all the types.

The interoperability of metadata i.e. the ability of different information systems to inter-operate or be compatible with each other's vocabularies is seen as a fundamentally important issue in the development of Web-based information systems (Gill and Miller, 2002). This is because it is valuable if two (or more) systems holding data on similar things can be reliably cross searched and/or share data. Many standards, initiatives and projects are in place to develop systems that will be able to interoperate at a vocabulary and semantic (meaning) level (e.g. W3C 2002b, Miller 2001, see also below).

Part of the development of interoperable Web-based systems includes the creation of systems that utilize semantically interoperable ways of describing things, characteristics of things, and the relationships between them. These 'ontologies' (e.g. Ontologies W3C initiative, W3C 2002b) take the form of structured machine readable representation of the knowledge.

> Just like people need to have agreement on the meanings of the words they employ in their communication, computers need mechanisms for agreeing on the meanings of terms in order to communicate effectively. Formal descriptions of terms in a certain area (shopping or manufacturing, for example) are called ontologies and are a necessary part of the Semantic Web. RDF [Resource Description Framework], ontologies, and the representation of meaning so that computers can help people do work are all topics of the Semantic Web Activity.
> *(W3C 2001b)*

These developments form what can be seen as part of a larger movement in Web technology development towards a more semantically interoperable Web (W3C 2001a, Berners-Lee et al 2001) in which information is globally interoperable.

There are significant difficulties with building ontologies and applying them to bodies of information. Ontology creation and application is a very specialized and time-consuming activity. Even more difficult is mapping between ontologies, especially those written by different communities of interest. An ontology provides a machine processable hierarchy of terms, but not all of the intentions of the ontology creator are encoded into the description of the ontology. Therefore mapping between them is prone to errors of interpretation.

Part of a Solution: Community of Interest/Expertise Annotation

The development of more semantically interoperable Web-based technologies seems to promise the ability to solve part of the challenge outlined above; namely, that of enabling machines (computers) to relate terms from different specialist vocabularies about what are essentially the same thing or concept and thus being able to map existing terms to the specialist vocabularies (including other languages).

However they do not provide a solution to the problem that members of specialist interest groups will want to describe (apply metadata to) data in ways relating to totally different concepts.

A simple example makes the issue clearer:

Imagine that there is a database of 100,000 images of people in a wide variety of different settings, say developed for a news agency. The database may be indexed using terms relating to identification of people (name, age, ...) and event (time, place...) as well as administrative, preservation and technical metadata. This is because those are the important characteristics to those who originally setup the database.

Now milliners might see great potential for studying how people use hats. The database is likely to be a very useful resource:, they could search to see how the style of hats has changed over time, or what types of hats are most popular; they could answer many more specific questions, e.g. what percentage of women have bows on their hats? or wear a particular type hat at a particular time of year?... however the database is not indexed using the concept of 'hat' and so it is not possible to interrogate it to find the answers to these questions.

Imagine now that someone else, a landscape architect, comes across the database and sees that it could be used to study how public seating is used in urban settings...

In each of these examples the collection *could be* of very great value to the user, but the existing indexing was not originally designed with these uses in mind and so it is not. It is in these cases that *community annotation* of a collection, could offer the key to meeting these needs and thus greatly extend the scope and value of an on-line collection.

In the example above the individuals are from communities of 'milliners' and 'landscape architects'. They could annotate the images with specialist indexing terms used by their communities, ideally from ontologies developed to facilitate a semantically consistent representation. However it might be that they simply want to add 'notes' to particular images for others from their communities to find or make a hyper-link to a page of more detailed complementary information or case studies of that kind of example.

Models of Community Annotation

This section outlines a number of models of community annotation that we have identified, and that we believe help meet the diverse needs of different 'user communities' and the database system developers such as ARKive.

Figure 1 shows via a much simplified Venn diagram some of the communities of interest of an ARKive type project. There are the more traditional 'target users', in ARKive's case, those with interests in biodiversity and wildlife media and their sub-communities A1, A2, A3... (e.g. different sub-disciplines, phases of education...), and ARKive's own staff. However there are other 'communities of interest' (B, C, D and E) that lie outside or may have some degree of overlap with the original target community or communities.

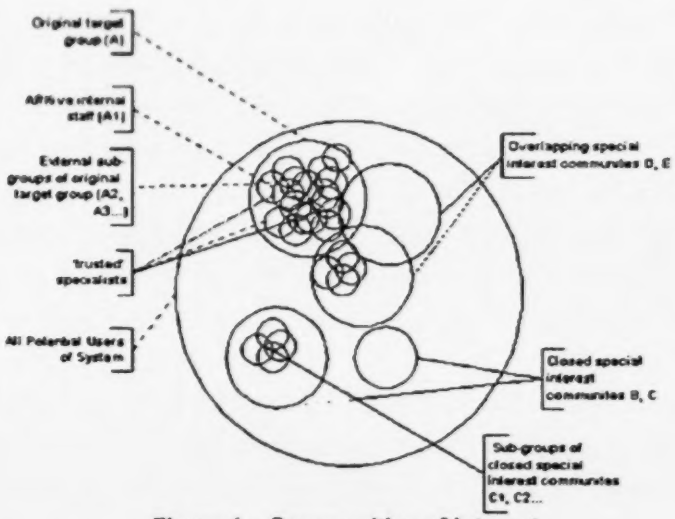
Original target group (A)
ARKive internal staff (A1)
External sub-groups of original target group (A2, A3...)
'trusted' specialists
All Potential Users of System
Overlapping special interest communities D, E
Closed special interest communities B, C
Sub-groups of closed special interest communities C1, C2...

Figure 1 – Communities of Interest

Whilst the discussion here is focused on the use of community annotation to apply 'specialist' indexing to objects, the majority of the issues discussed are the same for other kinds of annotation. Examples include case studies in the

BEST COPY AVAILABLE

use of a particular image or media type, or notes relating to the object (e.g. what it shows, interesting facts, controversies).

Possibly the most critical issues related to 'community annotation' for the organizations behind ARKive-type Web sites are related to the quality, accuracy and relevance of any annotation. One of ARKive's fundamental values is ensuring that it provides scientifically accurate and up-to-date information.

If users are given tools that enable them to add/link 'metadata' to multimedia assets, many issues arise about how that annotation can, or should be, made available to other users of the system. Below we have outlined four models of annotation that we have developed to illustrate the issues.

1. **Trusted members of trusted communities**: ARKive already has a group of 'trusted' experts and organizations that provide and validate information which ARKive collates and publishes on its Web site. This approach would expand this single group to wider communities; for example, a 'trusted organization' could provide a list of potential members that it states are competent to annotate ARKive resources. Likely the resources and 'metadata terms' that they are permitted (by ARKive) to annotate and use, would be limited to a specific set within their area of competency and that the types of annotation or vocabulary of concepts would also be limited (see consistency below).

   These 'trusted members' could be given usernames and passwords and on-line tools to annotate ARKive resources. These annotations would thus in principle be 'ARKive approved annotations' and thus in effect integral to the ARKive cataloguing and indexing system. However, it may be that only particular communities of users get access to particular sets of annotations.

2. **Self-selecting communities**: Much as there are self-selecting, open and closed discussion groups on academic mail lists such as JISCMail (http://www.jiscmail.ac.uk), it would be possible to provide annotation tools for a self-selecting and administering community of users.

   Such communities normally have 'community leaders' who administer them on a day-to-day basis. They vary greatly in nature from highly structured and constituted to loose with similar interests.

   It could be that only members of a 'closed' community get access to the annotations added by their group, or that these are made available to generic ARKive users but with a 'disclaimer'.

3. **Open annotation**: in this model any ARKive user could annotate an object. This is similar to ratings systems which exist on e-commerce sites such as Amazon.com (www.amazon.com). The degree of 'openness' may range from totally 'open' to a more 'mediated' approach in which some validation or quality criteria are applied (see 'Consistency and Quality Control' below).

   In this case, all annotation would have to have some form of 'disclaimer' as ARKive would have relatively little control over the quality, accuracy or relevance of any annotation.

4. **Third Party Annotation**: ARKive users might want to produce 'third party' sites to **draw** together resources from ARKive and other sites;

the simplest example of this would be a list of links to other Web sites. However, these third parties might also produce their own 'annotations' for ARKive resources; e.g. they might use a very subject-specific vocabulary. 'Annotating' resources from diverse sources has the advantage of enabling resource discovery across multiple information sources but with a particularspecialist vocabulary and personal control.

ARKive would not have control over this form of annotation but might want to provide infrastructures and tools to support such annotation.

Existing Projects

Many Web-based projects and sites use some form of annotation to 'add value' to their data. Each example below utilizes one of the approaches above:

- Amazon.co.uk (http://www.amazon.co.uk/) provides customers with the ability to add comments about a product and give it a star rating. This gives future users the 'added value' of hearing the views of others who had read, listened to, watched etc... the product. All users can see all annotations once they have been vetted for compliance with Amazon's guidelines. For guidelines see (http://www.amazon.co.uk/exec/obidos/subst/misc/author-review-guidelines.html/026-1836225-8318031)

- PseudoCAP: : Pseudomonas aeruginosa Community Annotation Project (http://www.cmdr.ubc.ca/bobh/PAAP.htm) allows Web-based annotation tools a to be used by members of a closed community. Since they aim toto:

- "... improve the quality of analysis of the *Pseudomonas aeruginosa* PAO1 genomic sequence, and to ensure the development and widespread availability of genetic tools to analyze Pseudomonas gene function, this *Pseudomonas aeruginosa* community annotation project (PseudoCAP) was initiated to enlist the expertise of volunteer Pseudomonas scientists in annotating the genome sequence. Annotations provided by the Pseudomonas scientists were subjected to peer review and used to aid the final genome annotation that was published. All participants in this project for the publication of the genome sequence have been acknowledged in the genome paper..."

- Slashdot (http://slashdot.org/) is an example of a 'community news portal.' It allows users to up-load news and then others can take part in a (threaded) discussion based on the original submission.

- Gimp-Savvy.com (http://www.gimp-savvy.com/) is a Community-Indexed Photo Archive which provides simple tools for users to add indexing terms to images in an on-line image database (http://gimp-savvy.com/PHOTO-ARCHIVE/)

- Berkman Center for Internet and Society (http://eon.law.harvard.edu/cite/annotate.cgi.) has developed an on-line tool (Annotation Engine) for users to annotate on-line documents by placing a link in the text at the point that the user wishes to annotate

- FishBase (http://www.fishbase.org/) is an interesting example of a

specialist site which is run and managed as a community. FishBase contributors have passworded access to edit and add data to the very extensive underlying database of scientific data and multimedia resources.

More generally the W3C are looking to develop 'annotation standards' under the Annotea project (W3C, 2002a) to allow users to collaboratively annotate Web pages. In parallel with these developments, the standards which support the use of metadata descriptions are expanding; e.g. the MPEG-7 standard (Martinez, 2001) for the 'content description' of multimedia includes a comprehensive 'Description Definition Language' which allows complex description of multimedia objects.

These projects all provide tools that enable users of different types to add value to the 'collections' by adding annotation.

Implementation of Community Annotation – Issues

Use of and Access to Annotation data

As can be seen above, the use of and access to any annotation is an issue inseparable from that of the under lying model of annotation. Probably the most fundamental issues are deciding who has access to any annotations and how any annotations (explicit, e.g. case studies, or implicit,e.g. search terms) are used and their use signaled. Some approaches follow.

1. **'Trusted community annotation'** might effectively be transparent to any users and simply be utilized as core ARKive tagging or alternatively might be provided via a specialist search engine option or provided with a disclaimer.

2. **'Self-selecting community annotation'** might be available to all users via a disclaimer or only available to 'subscribers' to that annotation (i.e. they explicitly request access) or only to members of the community.

3. **'Open annotation'** could be 'transparent' or have access controlled as in the previous two cases.

4. **'Third party annotation'** would mean ARKive would have no control over access to or use of this type of annotation. For user groups this might be seen as an advantage as the system is 'independent'.

Consistency and 'Quality Control'

Another fundamental issue with respect to using community annotation to assist with 'indexing' metadata is ensuring that there is consistency in both the terms used and the application of those terms, which relate back to precision and recall (see above).

The main solution to the issue of consistency has historically been to utilize a 'controlled vocabulary' with clear instructions about what those terms relate to; e.g. library catalogue systems. In particular, controlled vocabularies appear to improve consistency where indexing is being conducted by a number of 'indexers' (Markey,1984). In the case of ARKive there is already a controlled (bespoke) vocabulary for a number of aspects of the data and metadata used to describe the multimedia objects.

In order to annotate objects with relevant terms/concepts, it seems necessary to provide not only a controlled vocabulary but also a highly structured conceptual framework on which the vocabulary is based. This is because of the very large range of concepts that are covered by ARKive content, including bio- and bio-geographic sciences, wildlife film making, conservation and sustainable development, and educational uses.

These are broadly quality control issues; key questions for any system will relate to the degree of 'quality control' required for any annotations. This may extend from a check that annotations are not obscene or contravene legal requirements (e.g. libel laws) all the way to full multilevel verification by appropriate experts with formal 'sign off' of any new annotations.

The degree to which this is appropriate must depend on the nature of the system; e.g. a 'closed community' in which members of the community are the *only* users who can access the annotation may require no formal quality control (other than legal issues) from the collection/Web site owners. However 'trusted community' annotation that is accessible to all users may require significant quality control.

Tracking Annotation & Access Control – Annotation Metadata

If community annotation is used, there must be systems in place to manage the new data. That is the ability to track and maintain the annotations; it is necessary to have metadata about the annotations. Cross et al (2001) show how this kind of data can be created and maintained. The potential value of this data is significant, as it potentially enables users (internal to the organization or external) to query the system to say "show me all the annotation to the collection (or subset) made by person 'x' or members of community 'y'". This forms the basis of providing controlled access to the annotation data.

Extensibility

A further requirement of any system of annotation focused on adding value to a collection by 'indexing' is that it be extensible; i.e. that new terms can be added in a coherent and meaningful way.

For example when a new 'term' is added it must be done in such a way that concepts of which it is a sub-element (e.g. pecking might be a sub-set of feeding or defensive behaviors) retain conceptual integrity, e.g. the term 'pecking' should not be applied to an object that does not have a related parent concept (e.g. feeding) or that existing parent concept must (henceforth) be made to apply to the object as well. Hence there may be a need to create new non-overlapping sub-categories of pecking; e.g. feeding:pecking and defensive-behavior:pecking.

This simple example shows that the creation of any conceptual representation will be very problematic. However the current authors believe that without such a framework, effective use of annotation would be very problematic if not impossible to manage and monitor. Heflin and Hendler (2000) explore the complexities of making changes to formal ontologies and some the many associated problems.

Other issues include how to deal with and represent 'controversial knowledge', 'fallacies', 'old knowledge' and other forms of 'inconstancies' in any knowledge base. There are no simple answers to these problems. Once

again the most appropriate solution will depend on the particular situation; e.g. in the case of relatively open annotation such as gimp-savvy.com (http://www.gimp-savvy.com/) it may be appropriate for any user to be able to add indexing terms (given legal considerations are dealt with, see above) whilst in the case of 'trusted community' annotation, changes to the ontology or vocabulary used might require a formal meeting of some form of 'expert panel'.

However whatever form it takes, we argue that there must be some system (s) to facilitate such extension of the available terms and concepts if the overall systems are to be effective and sustainable.

A further fundamental problem is that community annotation using extensible annotation vocabularies and schemes is post-hoc and thus, unless every object is systematically annotated, it is very likely that the some objects will not be tagged with a new type of annotation e.g. a particular indexing term, when it 'should' be. Thus the annotation or indexing becomes inconsistent across the collection. There seems to be no simple solution to this problem other than systematic annotation. However as outlined in the next section, the use of 'semantically aware' tools may provide a means of optimizing the completeness of annotations across a collection where (as in most cases) there are time and resource constraints.

Semantic Bootstrapping

One very interesting requirement that we have identified for all of the models described above is 'semantic bootstrapping'; that is, when a collection has been 'indexed' from one perspective (i.e. for its primary target use or user group) using one set of vocabularies, it is necessary to have or create some kind of 'semantic hook(s)' in the data to allow users to begin the process of indexing the collection from the new perspective, using the new vocabulary.

This is a form of 'semantic bootstrapping', conceptually related to the idea developed by Pinker (1984) to refer to his postulated process by which children 'semantically bootstrap' or learn syntax from some form(s) of built in 'semantic categories' and contexts.

Ontology-based tools (see above) could allow existing ontologies to be linked to the already-present vocabularies or ontologies via concepts common to both existing and new domains.

Concept extraction tools such as the 'Non Zero Match' tool were developed at the University of Bristol (http://nzm.dig.bris.ac.uk/index.html). The tool allows users to auto-index text-based documents using concepts defined by a list of words/phrases with positive and negative weights. E.g. say a 'car' by defining the concept via the occurrence of a set of words or phrases 'registration number', 'steering wheel', 'make', 'model' etc... the parser then processes the whole corpus of documents indexing the documents under the appropriate concepts. Thus by using existing text or indexing/markup it would be possible to create new concepts to help 'bootstrap' the new indexing.

Another example is described by Bobrovnikoff (2000) using the DIPRE (Dual Interative Pattern Relation Extraction) algorithm, to recognize pattern in existing data. Auto-indexing of still and moving images also provides the potential to extract and index new concepts; e.g. in the example above of looking for 'hats' in the database of images of people. See Campbell et al

(1997) and Lew (2000) for examples of this approach.

There are various forms that semantic bootstrapping could take, with various levels of automation. It could be a time-consuming and highly skilled manual task, effectively re-indexing the database manually by re-cataloguing by placing the images within an ontology used by a community of interest, or by using a controlled vocabulary. At the opposite end of the spectrum, the images could be auto-classified using specialist tools for pattern recognition.

Somewhere in between is a stored search by a subject expert. For example, when people search Google for a particular topic, they use their knowledge of their subject area and their common sense coupled with their experience of the content of the Google database itself to choose search terms that will accurately retrieve the information they require. For example someone looking for 'flying things' would use more specific search terms like 'bird', 'helicopter' 'parrot'.

An annotated stored query of a database by someone with knowledge of the indexing terms used in the database and the specialist subject knowledge of the community of interest would enhance the value of the database to that community. Such an annotation would provide fast approximate information for that community. An example might be a search of a database for photos of people dressed for a 'formal event' to get pictures of hats.

If there were time, the user could go through the images found in this way to check if the retrieved pictures were in fact pictures with hats, discarding those that were not. However, even a quick annotated search could provide added value.

Some form of 'semantic bootstrapping' will be essential in making annotation work effectively for communities. Different types of 'semantic bootstrapping' tools are likely to assist with different types of problem, and hence it is likely that what is needed is a suite of tools rather than one single tool or approach.

Moving Forward

We are working to formalize the models outlined above and are developing more detailed technical requirements for the implementation of the models. The ideal is that we design an approach that can allow ARKive type organizations to implement any and all of the models of annotation outlined above.

In parallel we are investigating the advantages and disadvantages of the different models in different contexts in order to help developers make decisions about which ones are the most appropriate for their particular needs and contexts. Parallel investigation into different approaches to 'semantic bootstrapping' and development of appropriate tool sets will continue.

Conclusions

Community annotation offers the developers of large multimedia database systems the ability to support specialist 'communities of Interest' and thus enhance the value of their data. There are many technologies available and under development that would support this approach; some projects are already utilizing them.

This paper has dealt primarily with the annotation of multimedia objects with specialist indexing/resource discovery terms and the associated technologies; however, the issues are similar for more generic types of annotation.

The four models of community annotation outlined in the paper provide a framework for the development of community-based approaches to enhance the value of Web-based museum and multimedia collections for specialist communities of interest.

There are many implementation issues that remain highly problematic, in particular the coherent and consistent extensibility of vocabularies and the development of 'semantic bootstrapping' tools.

However, it will likely be possible, in the short to medium term, to find solutions by assessing needs and matching solutions in each specific case. In the longer term, we hope that the on-going development of a more 'semantically interoperable' Web and associated technologies will lead to the creation of sets of approaches and tools to make the implementation of community-based annotation relatively simple and effective.

Acknowledgements

The ARKive-ERA project is funded by HP-Labs, Bristol. The authors wish to thank Dave Reynolds of HP Labs, for discussing, exploring and expanding the ideas related in this paper.

References

Berners-Lee, T., Hendler, J. and Lassila, O. (2001) The Semantic Web, Scientific American, May 2001.

Buckingham Shum, S., E. Motta, et al. (2000). International Journal on Digital Libraries **3**(3): 237-248.

Chowdhury, G. G. (1999) Introduction to Modern Information Retrieval, Library Association Publishing, London.

Clark, J. (2001). "Subject portals." Ariadne(29). Available on-line: http://www.ariadne.ac.uk/issue29/clark/

Cross, P., Miller, L., and Palmer, S. (2001). Using RDF to Annotate the (Semantic) Web. K-CAP Workshop Knowledge Markup & Semantic Annotation, Victoria B.C., Canada.

DELESE (2001). Digital Library for Earth System Education (DELESE) Web site. http://www.dlese.org/

Eakins, J. and M. Graham (1999). Content-based Image Retrieval, JTAP (Joint Technology Applications Programme).

European Environment Agency (2002). GEneral Multilingual Environmental Thesaurus (GEMET): The GEMET 2.0 Approach. Available on-line: http://www.mu.niedersachsen.de/cds/etc-cds_neu/library/Gemet.pdf

Gill, T. and P. Miller (2002). "Re-inventing the Wheel? StanJards, Interoperability and Digital Cultural Content." D-Lib Magazine 8(1).

Gilliland-Swetland, Anne J. "Setting the Stage: Defining Metadata" in Introduction to Metadata: Pathways to Digital Information, Murtha Baca, ed. (Los Angeles: Getty Information Institute, 1998) Available on-line: http://www.getty.edu/research/institute/standards/intrometadata/2_articles/index.html

Heflin, J. and J. Hendler (2000). Dynamic ontologies on the Web. Seventeenth National Conference on Artificial Intelligence (AAAI-2000).

Lew, Michael (2000) Next-Generation Web Searches for Visual Content, IEEE Computer, 33(11) p46-53, November, 2000

Markey, K. (1984) "Interindexer Consistency Tests: A Literature Review and Report of a Test of Consistency in Indexing Visual Materials.", Library and Information Science Research, 6, 155-177.

Martinez, J. M. (2001). Overview of the MPEG-7 Standard (version 6.0), MPEG (Moving Picture Experts Group). Available on-line http://mpeg.telecomitalialab.com/standards/mpeg-7/mpeg-7.htm

Miller, P. (2001) Interoperability Focus homepage. Available on-line http://www.ukoln.ac.uk/interop-focus/.

Campbell, N. W. Mackeown, W. P. J., Thomas, B. T, and Troscianko, T. Interpreting Image Databases by Region Classification, Pattern Recognition (Special Edition on Image Databases), 30(4):555-563, April 1997.

Qualifications and Curriculum Authority (2002). National Curriculum Online Metadata Standard Overview, Available on-line http://www.nc.uk.net/metadata/

Schreiber, A. T., B. Dubbeldam, et al. (2001). "Ontology-Based Photo Annotation." IEEE Inteligent Systems May/June 2001: 2-10.

W3C. (2001a) Semantic Web. Available online: http://www.w3.org/2001/sw/.

W3C (2001b), XML-in-10-points, Available online: http://www.w3.org/XML/1999/XML-in-10-points/

W3C. (2002a) Annotea Project Homepage. W3C. Available: http://www.w3.org/2001/Annotea/.

W3C. (2002b) Web-Ontology (WebOnt) Working Group Homepage. W3C. Available: http://www.w3.org/2001/sw/WebOnt/.

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

NOTICE

Reproduction Basis

X	This document is covered by a signed "Reproduction Release (Blanket)" form (on file within the ERIC system), encompassing all or classes of documents from its source organization and, therefore, does not require a "Specific Document" Release form.
	This document is Federally-funded, or carries its own permission to reproduce, or is otherwise in the public domain and, therefore, may be reproduced by ERIC without a signed Reproduction Release form (either "Specific Document" or "Blanket").